Predicting PCOS Diagnosis from Clinical Data

Jane Zulu

BANA 7365

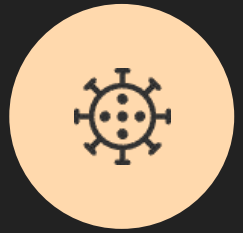
Predictive Modelling

April 13, 2025

Executive Summary

Polycystic Ovary Syndrome (PCOS)

A hormonal disorder affecting 1 in 10 women of reproductive age, often overlooked due to vague or inconsistent symptoms.



Accurate Diagnosis is Challenging

Current methods rely on manual interpretation of symptoms and lab values — leading to misdiagnosis, treatment delays, or missed cases.

Need for Accessible Tool

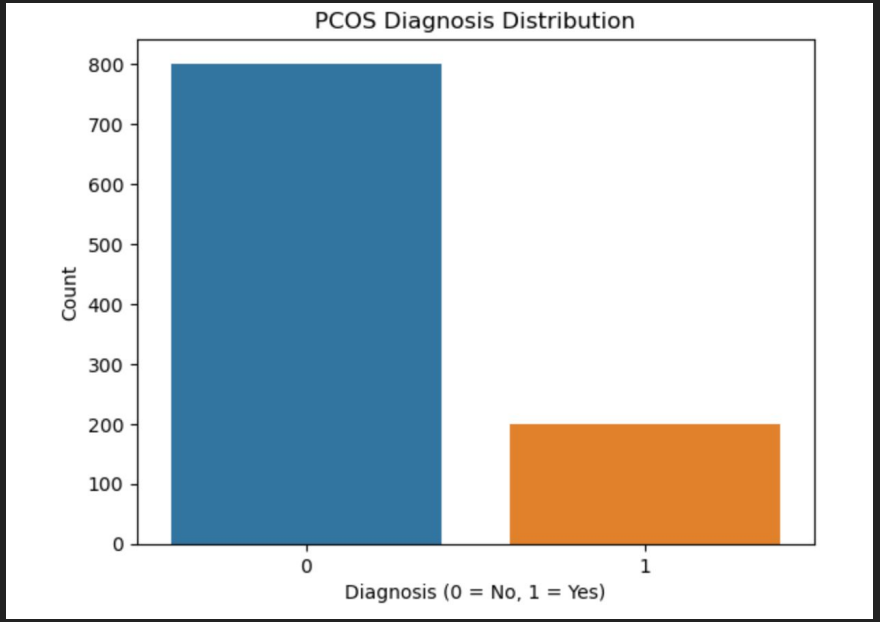
By applying clinical data, we can support clinicians with a faster, more objective way to identify PCOS — even in early or unclear cases.

This project builds a data-driven model that improves PCOS diagnosis by combining medical insights with machine learning — enabling earlier detection and expanding access to care.

Dataset Overview

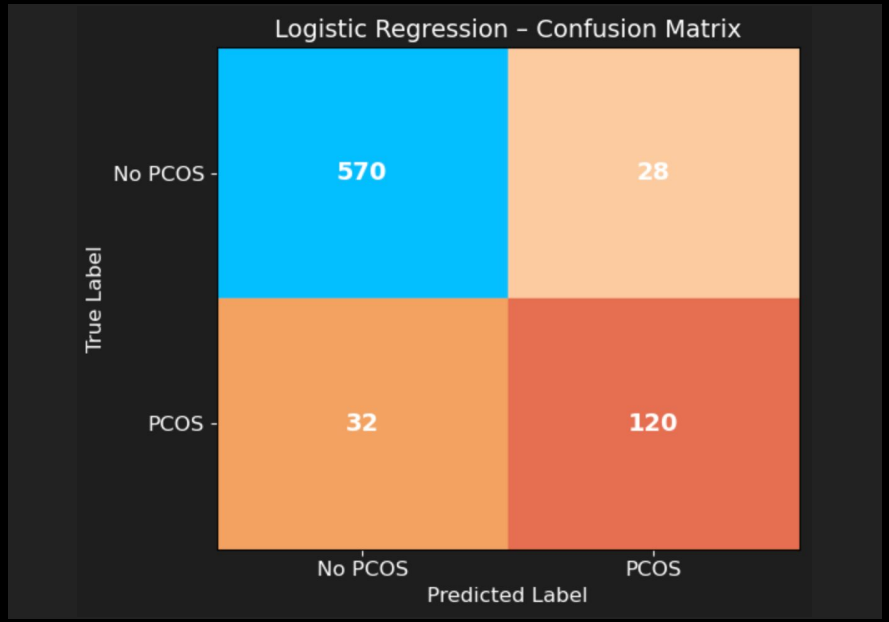
Clean and structured clinical data from 1,000 patients was used to train the model. Despite class imbalance, key variables showed meaningful differentiation between PCOS and non-PCOS cases.

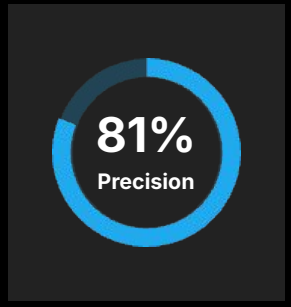
- **Dataset Summary**
  - 1,000 patient records
  - 6 clinical variables
  - 80% No PCOS, 20% PCOS — **imbalanced but realistic**
  - **https://www.kaggle.com/datasets/samikshadalvi/pcos-diagnosis-dataset**
- **Key Health Indicators**
  - Age (18 - 45 years)
  - BMI (kg/m²)
  - Menstrual Irregularity (0 = Regular, 1 = Irregular)
  - Testosterone Level (ng/dL)
  - Antral Follicle Count (number)
- **Data Prepared for Modeling**
  - No missing values
  - All variables numeric
  - Scaled and ready for machine learning



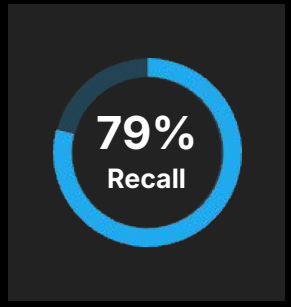
Baseline Model - Logistic Regression

This baseline model already shows strong performance, identifying 120 of 152 PCOS cases — but still misses 32, motivating the need for improved models.

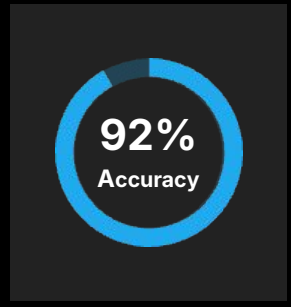




81% of predicted PCOS cases were correct (based on 148 predicted PCOS cases)



79% of actual PCOS cases were identified (120 out of 152 cases correctly identified)



92% overall accuracy (on 750 training patients via cross-validation)

Model Improvements

Model performance improved by validating features, testing algorithms, and optimizing parameters.

Validated Existing Features

Confirmed the clinical relevance of key variables already included in the dataset — such as BMI, testosterone level, and menstrual irregularity.

Explored Multiple Modeling Approaches

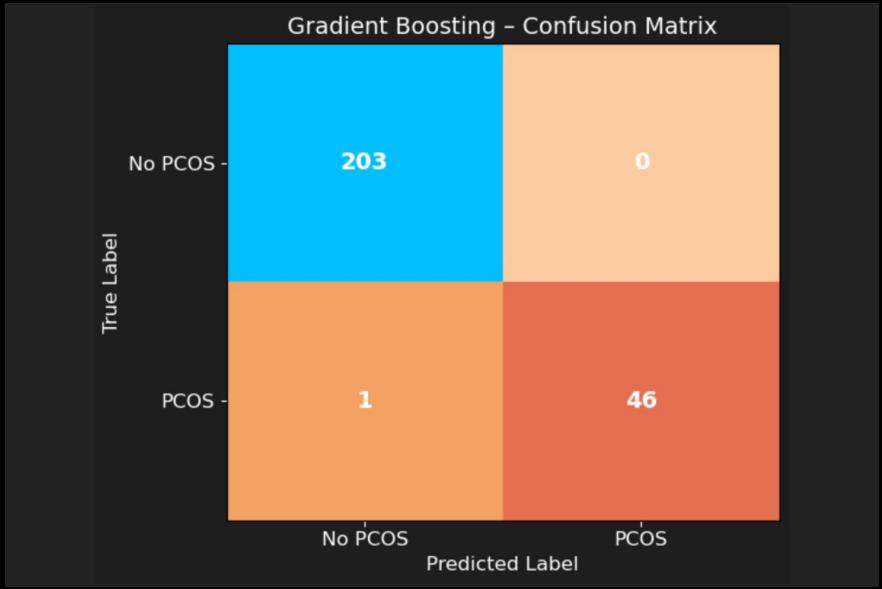
Tested several machine learning algorithms — including Random Forests, Gradient Boosting, and Support Vector Classifier (SVC) — to identify the most effective model for predicting PCOS.

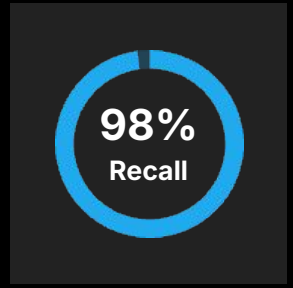
Fine-Tuned Model Settings

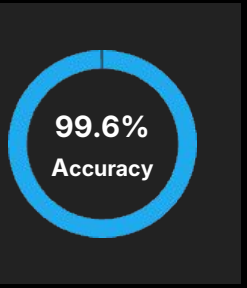
Optimized hyperparameters using cross-validation to improve accuracy and avoid overfitting — ensuring strong performance on new, unseen patients.

Final Model - Gradient Boosting

This Gradient Boosting model achieved near-perfect results on the test set. It correctly identified 46 of 47 PCOS cases and all 203 non-PCOS cases, reducing false negatives from 32 (baseline) to just 1. The performance boost highlights the model's ability to capture subtle patterns in clinical data.







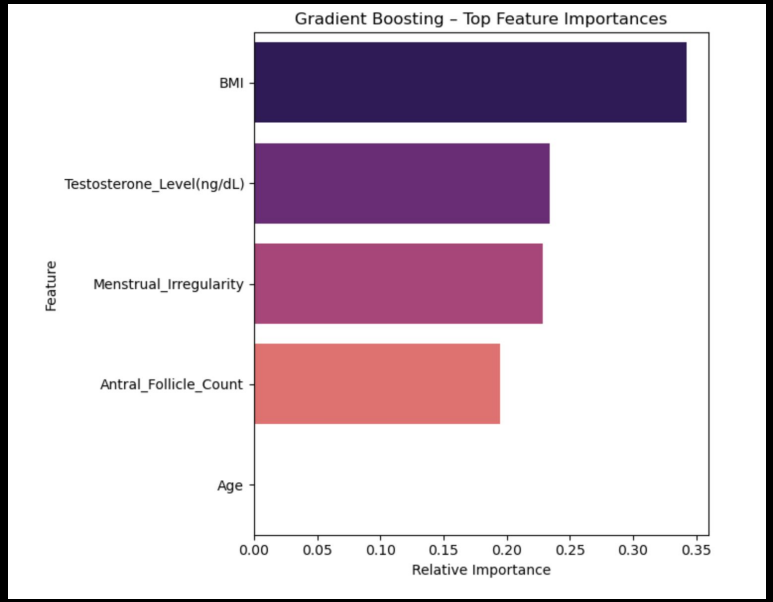
Every predicted PCOS case was correct (no false positives)

The model correctly identified 46 out of 47 actual PCOS cases

99.6% overall accuracy - 249 of 250 total cases correctly classified

Feature Importance

The model’s near-perfect accuracy stems from clear, medically validated indicators working together. Just four features account for over 98% of its decision-making power



BMI - 34%

Strongest predictor. Higher BMI is commonly linked with PCOS due to metabolic changes.

Testosterone Level - 24%

Hormonal imbalance — a core diagnostic factor.

Menstrual Irregularity - 22%

Classic symptom reflecting reproductive disruption

Antral Follicle Count - 18%

Ultrasound-based clinical indicator of PCOS.

Conclusion

Key Findings

Developed a predictive model that can accurately identify Polycystic Ovary Syndrome (PCOS) from clinical data

Potential Impact

Provides a cost-effective, accessible tool to support earlier diagnosis and better outcomes for patients.

Model Performance

The final model achieved **99.6% accuracy**, correctly identifying 46 out of 47 PCOS cases.

Next Steps

Explore model refinement and evaluate its effectiveness in real-world clinical settings.

In summary, the developed predictive model demonstrates the potential to transform PCOS diagnosis by leveraging clinical data, offering a reliable and accessible tool to healthcare providers. The positive results highlight the model's ability to enhance early detection and management of this complex condition, ultimately improving the lives of those affected by PCOS.

THANK YOU